Publications

Purchase PDF - $49.00Publication can be printed, downloaded but can not be redis tributed.Buy Now cover

Accelerating Real-Time Speech Recognition with OpenAI Whisper and Metal Performa nce Shaders on macOS

Sebastien Rousseau - 12 March 2024 - English(95 KB - PDF)

AbstractReal-time automatic speech recognition (ASR) has become an increasingly crucial technology in today's world, enabling a wide range of applications such as live transcription, voice assistants, and dictation. However, performing high -quality ASR inference with low latency remains a computationally demanding task , often requiring powerful cloud servers. In this paper, we present a system for real-time speech-to-text transcription that leverages the OpenAI Whisper speech recognition model and accelerates inference on macOS devices using the Metal Per formance Shaders (MPS) GPU back-end. By utilising on-device GPU acceleration, ou r system significantly reduces the latency and computational requirements of rea l-time speech recognition while maintaining high accuracy. The system is impleme nted in Python and can easily be integrated into a variety of applications requi ring live speech-to-text capabilities. We evaluated the performance of our syste m on a M1 Max MacBook Pro and demonstrated that it achieves sub-second latency a nd 8-12 times faster than real-time transcription for typical utterance lengths of 5-10 seconds. Our energy-based voice activity detection approach achieves 94% precision and 96% recall in detecting speech segments. The proposed system showc ases the potential of leveraging state-of-the-art ASR models and GPU acceleratio n for efficient and accurate real-time transcription on edge devices, making it an ideal solution for applications requiring live speech-to-text capabilities on

macOS devices.

divider